

Omid Kokabi, Fabian Brinkmann, Stefan Weinzierl

Prediction of speech intelligibility using pseudo-binaural room impulse responses

Journal article | Accepted manuscript (Postprint)

This version is available at <https://doi.org/10.14279/depositonce-9006>



The following article appeared in

Kokabi, O., Brinkmann, F., & Weinzierl, S. (2019). Prediction of speech intelligibility using pseudo-binaural room impulse responses. *The Journal of the Acoustical Society of America*, 145(4), EL329-EL333.

and may be found at

<https://doi.org/10.1121/1.5099169>

Terms of Use

Copyright (2018) Acoustical Society of America. This article may be downloaded for personal use only. Any other use requires prior permission of the author and the Acoustical Society of America.

WISSEN IM ZENTRUM
UNIVERSITÄTSBIBLIOTHEK

Technische
Universität
Berlin

Prediction of speech intelligibility using pseudo-binaural room impulse responses

Omid Kokabi,^{a)} Fabian Brinkmann, and Stefan Weinzierl

Audio Communication Group, Technical University of Berlin, Einsteinufer 17c, 10587
Berlin, Germany

kokabi@tu-berlin.de, fabian.brinkmann@tu-berlin.de, stefan.weinzierl@tu-berlin.de

Abstract: Head orientation (HO) affects better-ear-listening and spatial-release-from-masking, which are two key aspects in binaural speech intelligibility. To incorporate HO in speech intelligibility prediction, binaural room impulse responses (BRIRs) for every HO of interest could be used. Due to the limited spectral bandwidth of speech, however, approximate representations might be sufficient, which can be measured more quickly. A comparison was done between pseudo-BRIRs generated with a motion tracked binaural microphone array and a first order Ambisonics microphone using the spatial decomposition method (SDM). The accuracy of the Ambisonics/SDM approach was comparable to that of real BRIRs, indicating its suitability for speech intelligibility prediction.

1. Introduction

Better-ear listening and binaural unmasking are the most important aspects in binaural speech intelligibility (Middlebrooks *et al.*, 2017). Better-ear listening refers to the fact that the auditory system primarily extracts information from the ear signal with the more favorable signal-to-noise ratio (SNR). Binaural unmasking refers to the reduction of the masking effect of a competing source on a speech target when the two are spatially separated (Kock, 1950). Head orientation (HO), i.e., the listener choosing a favourable viewing direction, can significantly affect both phenomena, even without head movement induced dynamic cues. Consequently, an improvement of the speech reception threshold of up to 8 dB was observed as a result of an optimized HO in an anechoic environment (head orientation benefit, HOB; Grange, 2016). In a more realistic restaurant environment with moderate reverberation, an HOB of about 3 dB was observed by Grange and Culling (2016). Both perceptual aspects can be well estimated with existing binaural models [e.g., Jelfs *et al.* (2011)]. The model input is typically either a binaural stream of the speech and masker signal at both ears, or a binaural room impulse response (BRIR) that describes the transfer path from the speech and masker sources to the binaural receiver. To incorporate HO in modeling speech intelligibility, BRIRs are required for every HO of interest, and simulating or measuring these data is tedious and time consuming. BRIRs can also be calculated for different HOs from a set of multichannel room impulse responses (RIRs) captured at a single receiver orientation with a spherical microphone array. However, a high spatial resolution is required to avoid errors in the captured sound field representation for the entire audible frequency range. For example, a spherical harmonic (SH) decomposition of head related impulse responses (HRIRs = BRIRs under anechoic conditions), an SH series of up to order 35 is required to synthesize HRIRs without perceptual artifacts, so that a spherical microphone with around 1300 sensors would be required for a one-shot measurement (Bernschütz, 2016). However, because better ear listening and binaural unmasking mainly rely on interaural time differences and interaural level differences (ITDs/ILDs), and speech contains relevant energy only below 10 kHz, a simplified sound field representation might be sufficient in the context of speech intelligibility modeling. In the present work, approximated BRIRs (termed *Pseudo-BRIRs* in the following) were calculated for arbitrary HOs in the horizontal plane (i.e., for head rotations to the left and right) from multichannel RIRs captured with (a) a motion tracked binaural (MTB) microphone array and (b) a first order Ambisonics microphone. The prediction accuracy of the Pseudo-BRIRs was compared to a set of

^{a)}Author to whom correspondence should be addressed.

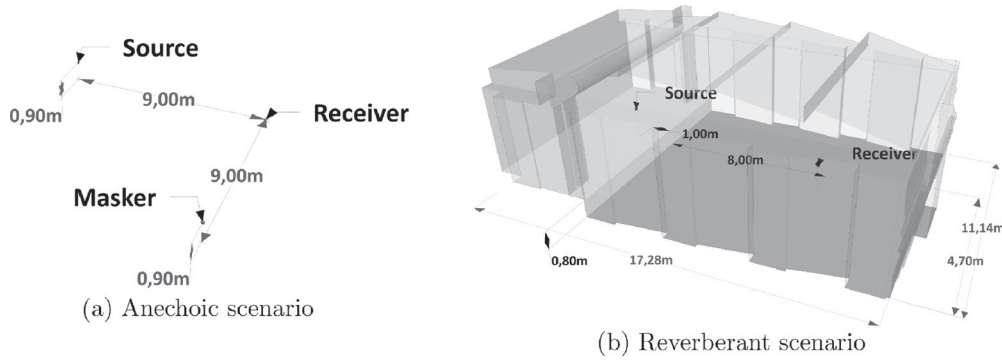


Fig. 1. Sketch of the evaluated acoustic scenarios.

reference BRIRs simulated for *each* HO. In both cases an already existing and validated binaural model was used.

2. Methods

2.1 Acoustic setup and stimuli generation

BRIRs and Pseudo-BRIRs were simulated using the geometrical acoustics simulation software RAVEN (Schröder and Vorländer, 2011) for an anechoic environment with a frontal speech target and a lateral masking source, and a reverberant room with a frontal speech target and late reverberation considered as masker (cf. Fig. 1). The reverberant room was modeled with two different mean reverberation times of $T_{20,m} = \{0.5, 1\}$ s (averaged across one source and three receiver positions) while maintaining a realistic frequency dependence of the applied absorption coefficients. The direct-to-reverberant energy ratio D/R , calculated as the energy ratio of the direct to reverberant part from a RIR of an omnidirectional receiver with a time limit of 2.5 ms to separate the two parts, was calculated as -1.1 dB ($T_{20,m} = 0.5$ s) and -6.6 dB ($T_{20,m} = 1$ s) at the receiver position shown in Fig. 1. For the target and masker source, the directivity of a male singer was taken from RAVEN (average directivity factor Q of 1.5 for 500 Hz and 1 kHz octaves), while head-related transfer functions (HRTFs) of the FABIAN head-and-torso simulator, measured in the anechoic chamber of the Carl von Ossietzky University Oldenburg (Brinkmann *et al.*, 2017a; Brinkmann *et al.*, 2017b) were used as receiver directivity for the BRIRs. The Pseudo-BRIRs were calculated based on two different receiver directivities: A MTB microphone with 16 equally distributed sensors on the equator of a rigid sphere with a diameter of 8.75 cm (Algazi *et al.*, 2004), and an idealized first order Ambisonics microphone consisting of one omnidirectional and three figure-of-eight sensors pointing to the front, the left, and the top, respectively. With the MTB, Pseudo-BRIRs were obtained from signals on opposite sides of the sphere, and intermediate positions were obtained from two-band spectral interpolation restoration (Algazi *et al.*, 2004). This method performs a linear interpolation of the time signals for frequencies below 2.5 kHz, whereas the magnitude spectra are interpolated above 2.5 kHz and combined with the phase spectrum of the closest microphone. The directivity of the MTB was measured in the anechoic chamber of Technische Universität Berlin with the system described in Fuß *et al.* (2015). Using the Ambisonics microphone, Pseudo-BRIRs were calculated with the spatial decomposition method (SDM) (Tervo, 2016; Tervo *et al.*, 2013). SDM estimates the direction and time of arrival of direct and reflected sound, and obtains the Pseudo-BRIR by superposition of all sound events, weighted by the HRTFs (taken from the FABIAN HRTF database) that correspond to the respective events. In all cases, BRIRs were calculated for 360 HOs in the horizontal plane between $\pm 180^\circ$ in steps of 1° . A detailed list of the settings used for the simulation with RAVEN and the calculation of the MTB/SDM-BRIRs is given in Kokabi *et al.* (2018a) along with the room models and BRIRs as .wav-files.

2.2 Prediction of speech intelligibility

The generated BRIRs and Pseudo-BRIRs were applied to the Cardiff binaural intelligibility model (Jelfs *et al.*, 2011) which is part of the auditory modeling toolbox (Søndergaard and Majdak, 2013). To account for the deteriorating effect of reverberation on speech intelligibility, the Cardiff model was extended by a room dependent time limit that splits the BRIR into an early, useful part considered as the target, and a late, detrimental part considered as the masker. The U/D (Useful/ Detrimental)-time

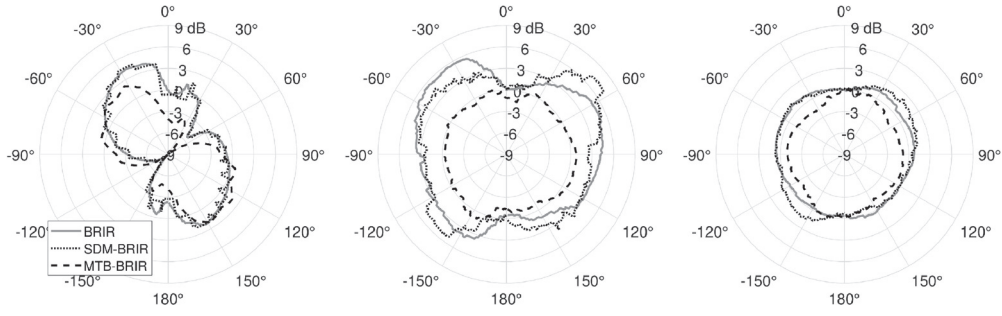


Fig. 2. Binaural benefit in dB for BRIR, Pseudo-BRIRs, and all evaluated HOs (positive angles depict HO to the right). Left: Anechoic scenario, speech target at 0° azimuth, masking source at -90° azimuth. Middle and right: Low and medium reverberant scenario ($T_{20,m} = \{0.5, 1.0\}$ s), speech target at 0° azimuth, masker = late part of BRIR).

limit was calculated from the interaural cross-correlation at the position of the receiver as shown in [Kokabi et al. \(2018b\)](#). The model output is a SNR in dB predicting the benefit of binaural listening over listening to an omnidirectional microphone at the same position. To get a clearer display of the prediction accuracy of HOB with the different approaches, the model output for the BRIRs and Pseudo-BRIRs are referenced to the model output of the BRIR facing the speech target ($=0^\circ$) by subtracting the latter from the former.

3. Results and discussion

The binaural benefit calculated with the Cardiff model as a function of HO with BRIRs and Pseudo-BRIRs is shown in Fig. 2 for the three different acoustic scenarios. HOs to the right are depicted with positive angles. For the anechoic condition (left panel) a distinct pattern of the binaural benefit as a function of HO can be observed for all conditions, illustrating improved better-ear listening and binaural unmasking with optimized HOs ([Grange, 2016](#)). Largest HOBs are found at about -30° and 150° azimuth with up to 5 dB compared to the frontal orientation. At these HOs, the speech target at 0° azimuth and the masking source at -90° azimuth are optimally separated by the dummy head and the spherical array resulting in an improved SNR at the listeners right (HO: -30°) and left ear (HO: 150°), respectively. The lowest HOB can be observed at HOs of -135° and 45° azimuth. At these HOs, the speech target and masking source are located at equal angles to the respective ear resulting in a maximized effect of the masking source. While the predictions based on the SDM-BRIRs matches the reference almost perfectly, a slightly lower binaural benefit was predicted for the MTB-BRIRs which can be attributed to the missing effect of the outer ears that results in a reduction of the ILD in these cases (cf. middle panel in Fig. 3). ITDs primarily affecting binaural unmasking are about equal for BRIRs and Pseudo-BRIRs (cf. right panel in Fig. 3). The stepwise structure that can be observed for the SDM-BRIRs is caused by the spatial resolution of the binaural summation process and could be reduced by increasing the number of so-called secondary sources.

For the low reverberance condition with $T_{20,m} = 0.5$ s (cf. middle panel in Fig. 2) the pattern of HOBs is generally less pronounced than in the anechoic scenario.

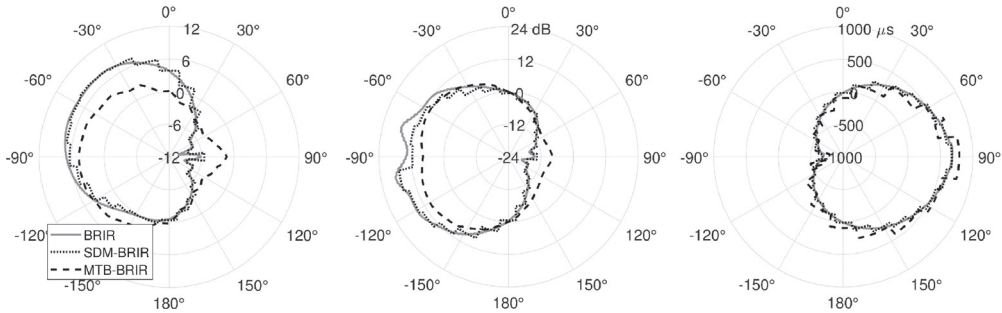


Fig. 3. Characteristics of HRIRs and Pseudo-HRIRs (=BRIRs/pseudo-BRIRs under anechoic conditions; 10 kHz lowpass filter applied). Angles depict sound source incidence (horizontal plane). Left: Logarithmic energy left ear signal, referenced to mean energy across all incident angles. Middle: Broadband ILD, estimated from logarithmic energetic differences between the left and right ear. Right: Broadband ITD estimated from threshold based onset detection (threshold 6 dB).

Table 1. MAE and MaxAE (in brackets) between predictions based on BRIRs and Pseudo-BRIRs for all acoustic scenarios and HOs.

	Anechoic	$T_{20,m} = 0.5$ s	$T_{20,m} = 1$ s	\emptyset
SDM-BRIR	0.5 (1.8)	1.1 (3.4)	0.5 (1.5)	0.7 (2.2)
MTB-BRIR	1.9 (4.6)	3.3 (5.5)	1.3 (1.6)	2.3 (3.9)

This seems plausible, as there is no distinct single masking source at a specific azimuth angle but rather diffuse reflected masking energy arriving from all directions reducing the effect of both binaural unmasking and head shadowing, which is relevant for better ear listening. Still, HOBs up to 5 dB (compared to a frontal orientation) can be observed for HOs between -30° and -60° , as well as between 30° and 60° azimuth in the case of the BRIRs and SDM-BRIRs. Here, the SNR at the listeners left (HO: 30° to 60°) and right ear (HO: -30° to -60°), respectively, is improved due to the ears spatial sensitivity (cf. left panel in Fig. 3). The improved HOB at these angles cannot be observed for the MTB-BRIRs, which can again be attributed to the missing outer ears that cause the increase in sensitivity. In all cases, the lowest benefits can be observed at HOs of 0° and 180° azimuth.

For the medium reverberance scenario with $T_{20,m} = 1$ s (cf. right panel in Fig. 2), the pattern of HOBs is almost omnidirectional in all cases. Apparently, head shadowing and the ears' spatial sensitivity are no longer improving the SNRs at either of the two ears as a function of HO against the diffusely reflected masking energy.

For a combined scenario of reverberation and a masking source at a specific azimuth angle, a more pronounced pattern of HOBs is expected to be observed, as the effect of head shadowing on the SNRs at the listeners ears will be strengthened.

The resulting differences between predictions with the BRIRs and the Pseudo-BRIRs for all scenarios are listed in Table 1 in terms of the mean absolute error (MAE) and the maximum absolute error (MaxAE) in dB across all HOs in the horizontal plane. As discussed above, predictions for SDM-BRIRs are closer to predictions with BRIRs (MAE: 0.7 dB) compared to predictions with MTB-BRIRs (MAE: 2.3 dB). Differences in prediction error between the two types of Pseudo-BRIRs are largest where the relative importance of the outer ears on the SNRs is maximized (here: the low reverberance scenario). For both BRIRs and Pseudo-BRIRs, the HOB vanishes with increasing level of reverberation.

4. Conclusion

The suitability of two different kinds of Pseudo-BRIRs for predicting speech intelligibility as a function of HO was assessed. Both methods allow an estimation of the HOB for speech intelligibility for arbitrary HOs of a binaural listener from a single measurement with a multichannel microphone array, thus making the physical re-orientation of a dummy head unnecessary, which is a common method for capturing BRIRs with different HOs. While the MTB method can predict the HOB only to a limited extent, due to the missing contributions of the pinna, the improvements in intelligibility due to HO can be well reproduced with the SDM method, based on a first order Ambisonics impulse response. In the anechoic scenario with a competing masking source, the mean difference in predicted binaural benefit based on a true BRIR and an SDM-BRIR is 0.5 dB. In reverberant scenarios with reverberation considered as the masking source, the mean deviation in binaural benefit is 1.1 dB for $T_{20,m} = 0.5$ s and 0.5 dB for $T_{20,m} = 1$ s. Compared to a general HOB of up to 5 dB (low reverberant scenario, $T_{20,m} = 0.5$ s), this appears to be an acceptable concession, since the estimation can then be made with a single measurement and a standard type of microphone. The validity of the method under ecological conditions, including the error introduced by the non-ideal characteristics of commercial Ambisonics microphones, will be covered in an empirical study using measured impulse responses from different existing acoustic environments.

References and links

- Algazi, V. R., Duda, R. O., and Thompson, D. M. (2004). "Motion-tracked binaural sound," J. Audio Eng. Soc. **52**(11), 1142–1156.
- Bernschütz, B. (2016). "Microphone arrays and sound field decomposition for dynamic binaural recording," Doctoral thesis, TU Berlin.
- Brinkmann, F., Lindau, A., Weinzierl, S., Geissler, G., van de Par, S., Müller-Trapet, M., Opdam, R., and Vorländer, M. (2017a). "The FABIAN head-related transfer function data base," available at [10.14279/depositonce-5718.2](https://doi.org/10.14279/depositonce-5718.2).

- Brinkmann, F., Lindau, A., Weinzierl, S., Müller-Trapet, M., Opdam, R., and Vorländer, M. (2017b). "A high resolution and full-spherical head-related transfer function database for different head-above-torso orientations," *J. Audio Eng. Soc.* **65**(10), 841–848.
- Fuß, A., Brinkmann, F., Jürgensohn, T., and Weinzierl, S. (2015). "Ein vollsphärisches Multikanalmesssystem zur schnellen Erfassung räumlich hochaufgelöster, individueller kopfbezogener Übertragungsfunktionen" ("A full-spherical multi-channel measuring system for the rapid acquisition of spatially high-resolution, individual head-related transfer functions"), *Fortschritte der Akustik DAGA Nürnberg*, pp. 1114–1117.
- Grange, J. (2016). "The benefit of head orientation to speech intelligibility in noise," *J. Acoust. Soc. Am.* **139**(2), 703–712.
- Grange, J. A., and Culling, J. F. (2016). "Head orientation benefit to speech intelligibility in noise for cochlear implant users and in realistic listening conditions," *J. Acoust. Soc. Am.* **140**(6), 4061–4072.
- Jelfs, S., Culling, J. F., and Lavandier, M. (2011). "Revision and validation of a binaural model for speech intelligibility in noise," *Hear. Res.* **275**(1), 96–104.
- Kock, W. E. (1950). "Binaural localization and masking," *J. Acoust. Soc. Am.* **22**(6), 801–804.
- Kokabi, O., Brinkmann, F., and Weinzierl, S. (2018a). "Assessment of speech perception based on binaural room impulse responses," available at [dx.doi.org/10.14279/depositonce-6725.2](https://doi.org/10.14279/depositonce-6725.2).
- Kokabi, O., Brinkmann, F., and Weinzierl, S. (2018b). "Segmentation of binaural room impulse responses for speech intelligibility prediction," *J. Acoust. Soc. Am.* **144**(5), 2793–2800.
- Middlebrooks, J., Simon, J. Z., Popper, A. N., and Fay, R. R. (2017). *The Auditory System at the Cocktail Party* (Springer, Berlin).
- Schröder, D., and Vorländer, M. (2011). "RAVEN: A real-time framework for the auralization of interactive virtual environments," in *Forum Acusticum*, https://www2.ak.tu-berlin.de/akgroup/ak_pub/seacen/2011/Schroeder_2011b_P2_RAVEN_A_Real_Time_Framework.pdf (Last viewed April 16, 2019).
- Søndergaard, P., and Majdak, P. (2013). "The Auditory Modeling Toolbox," in *The Technology of Binaural Listening* (Springer, Berlin, Heidelberg), pp. 33–56.
- Tervo, S. (2016). "SDMtoolbox" <http://de.mathworks.com/matlabcentral/fileexchange/56663-sdmtoolbox> (Last viewed April 16, 2019).
- Tervo, S., Pätynen, J., Kuusinen, A., and Lokki, T. (2013). "Spatial decomposition method for room impulse responses," *J. Audio Eng. Soc.* **61**(1/2), 17–28.